Some Reflections on the Potential and Limitations of Deep Learning for Automated Music Generation

Luca Casini
Department of Computer Science
University of Bologna
Email: luca.casini4@studio.unibo.it

Gustavo Marfia
Department for Life Quality Studies
University of Bologna
Email: gustavo.marfia@unibo.it

Marco Roccetti
Department of Computer Science
University of Bologna
Email: marco.roccetti@unibo.it

***Abstract*—Deep Learning and Artificial Intelligence are slowly revolutionizing many fields of applications, having the potential to replace humans in a variety of tasks and jobs. Nevertheless, creativity has always been considered something inherently human: recent research shows that, however, this may not always be the case. From this standpoint, this paper focuses on music and on the recent advancements in deep learning applied to the generation of musical content. We argue that, while those models are able to produce results that could actually be considered music, the role of the human musician still remains preponderant in the production of a musical piece. We here reflect on such limitations, directing our efforts to imagining new tools and instruments that may allow to experience new forms of interaction while supporting novel processes of creativity and music production.**

I. Introduction

Computer systems are steadily entering all domains of human life (medical, transportation, etc., [1], [2], [3], [4], [5], [6]). Deep learning techniques have the potential to represent one of next steps, as they may allow machines to outperform human beings in many tasks. In fact, a future in which activities that are now carried out by human will be performed by machines is no longer difficult to imagine: it is legit to wonder which jobs will be taken over by AI. While manual, as well as repetitive office tasks, are often included among those where humans will be replaced by machines, those tasks that instead involve some degree of creativity are believed to be safe. This widely accepted idea has not stopped researchers from trying to create generative models which are trained on artistic content. The most recent results are indeed impressive, still lacking some determining factors to become ultimately convincing.

This paper is dedicated to music and to the current trends in deep learning and artificial intelligence applied to symbolic music generation. We are going to overview the models and techniques behind the latest works in this field, highlighting the strengths and weaknesses of their proposed architectures in Section II. We will here also dive into some of the many unresolved issues that are being investigated by different research groups. After that we will describe some of the projects that, to this date, are focusing on an interactive use of such models for the benefit of music composers and amateurs (Section III. We then provide insights regarding the commercial applicability of such techniques, describing a few of the startups that have born working in this domain. We finally conclude with Section V.

II. Deep Learning Models for Music Generation

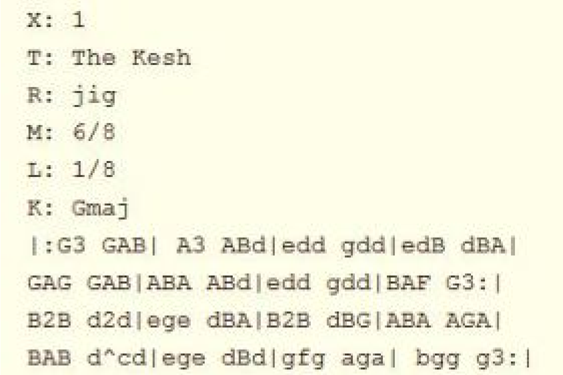
We are still far from a real AI composer: the human factor is still a fundamental part of the musical experience. In fact, as impressive as the state of the art models may appear, such models have so far accomplished to fool (untrained) listeners with very short pieces, while longer pieces reveal a lack of the structures and of the meaning that educated listeners would expect. In the following we proceed with their review, after providing some background required to understand how musical content is digitally represented.

A. Data Representation

The MIDI (Musical Instrument Device Interface) protocol is the standard way in which instrument communicate with computers and consists in a stream of events that carry information about notes such as pitch, attack and release (noteOn and noteOff) and volume (velocity) as well as control messages for the devices and metadata like time signatures and bpm. MIDI is cumbersome to work with for its event-based nature and for the lack of absolute time; after all this format was created to control digital instruments rather than store musical

978-1-5386-6009-6/18/$31.00 ©2018 IEEE

content. For this reason other kinds of representation emerged.



```
X: 1
T: The Kesh
R: jig
M: 6/8
L: 1/8
K: Gmaj
|:G3 GAB| A3 ABd|edd gdd|edB dBA|
GAG GAB|ABA ABd|edd gdd|BAF G3:|
B2B d2d|ege dBA|B2B dBG|ABA AGA|
BAB d^cd|ege dBd|gfg aga| bgg g3:|
```

Fig. 1. An example of abc notation

The ”abc” notation is a very simple format that represents melodies as text with pitches indicated by their letter name (A to G, ♯ and ♭ for alterations) and | to divide bars. This notation was introduced in the early 90s to share Irish folk tunes on the Internet and has since become the most popular format alternative to midi (it even has a MIME type). Often musician will say that ”music is a language”, and in the case of sequential formats like MIDI or abc notation this is definitely true as there is no conceptual difference between a sequence of words in a sentence and a sequence of notes in a song. Such a similarity gives the opportunity of applying the same models and techniques that are used for language modeling (predicting the next word) to music. Difficulties clearly arise from the fact that music is ”multidimensional” because multiple notes can be played at same instant, while natural languages are unidimensional as one cannot write or say more than one word simultaneously.

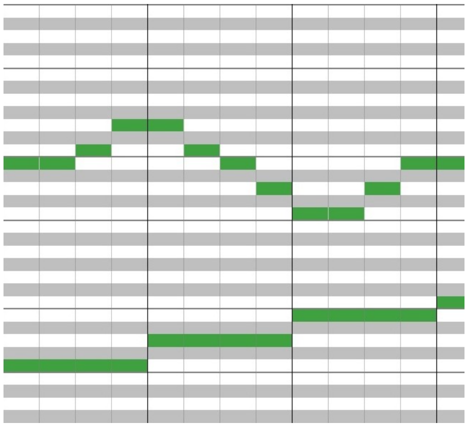


Fig. 2. An example of pianoroll notation

Another popular representation is the so-called Pianoroll, taking its name from the cylinders in old automatic player pianos. Pianorolls are matrices where each column corresponds to a time slice and each row to a specific note. This is basically the same as a black and white image, or gray-scale if note velocity is included. This similarity is the reason Convolutional Neural Networks, the most common and successful deep architecture for images, can be applied to music. However, while images are invariant across the two dimensions, music is not, as a different pitch (our x-axis) carries a different meaning, as arguably the same note or sequence when falling in different positions. It is also worth noting that while MIDI is a stream of events, pianorolls need to be quantized, introducing a trade-off between accuracy and complexity. Usually pianorolls are quantized to the 16th note for simple structures, to the 48th for more complex architectures. This quantization is good for predicting structure, but completely removes the possibility to learn the complex rhythms and timings of human performances.

B. Recurrent Neural Networks

Recurrent Neural Networks (RNN) are an architecture created to work with sequential data: the basic idea is to have each neuron in the network receive as its input both the current step of the sequence and the previous, carrying on those values across all time steps and allowing the network to model temporal dependencies. Long Short Time Memory(LSTM) cells solve the vanishing/exploding gradient problem of simple RNN introducing three gates (input, output and forget gates) to regulate the flow in each cell [7]. This architecture has demonstrated it’s effectiveness in language modeling [8], text generation and translation [9], as well as with image captioning [10] and generation [11].

As anticipated, given the similarities between natural languages and music, LSTMs have been used for music generation [12] [13]. The major drawback of such networks is that they can generate an output that is locally convincing but that lacks a higher order structure; this is true for both music and text. One way to overcome this is to impose a constraint on the structure of the music, both at the architectural level and in the dataset. Bach-styled Chorales generated in [14] and Irish folk music generated in [15] are an example of how a having a corpus with rigid structural rules makes it easier for the model to generate convincing results, even though expert ears can still identify them as artificial due to some dubious style decision or technical error. Nonetheless this shows how, with the right dataset and constraints, these kind of models can give results that are ”good enough” and only need a slight human tweak.

One advantage of LSTM is that, since they process data sequentially, they don’t need quantization and are therefore able to encode the subtle timing variations of a human performance. This possibility was only recently explored in [16], probably for the lack of MIDI files of human performances. The result is something that resembles a human pianist doodling at the keyboard, playing expressive yet aimless music.

C. *Variational Autoencoders*

The autoencoder architecture consists of two specular networks called encoder and decoder. The encoder reduces dimensionality to a small number of latent variables in the bottleneck layer and those latent variables are then used to reconstruct the input. This is basically a compression algorithm that learns meaningful features and stores them in the latent variables. Variational Auto-encoders (VAE) add the possibility of generation by allowing the model to sample from a Multivariate Gaussian Distribution to which the data are fit during the training process. This distribution can then be sampled quite easily and the obtained values may be fed to an encoder layer. VAEs have been used for image generation [17] but their output results blurry as a consequence of the compression step [18].

VAEs have been used for music generation in [19] where the authors also experimented with a stacked architecture. The results are certainly better than LSTM, although long term cohesion is still difficult and for this reason the authors limited themselves to a maximum of 16 bars. They also showed how VAE are capable of doing semantically relevant interpolation between two sequences. One drawback is the difficulty to manage harmony and polyphony given the enormous number of possible combinations: this may be the reason why the authors of [19] decided to only work with melodies.

D. *Generative Adversarial Networks*

Introduced by Ian Goodfellow in [20] GANs are a new paradigm that is based on an adversarial game between a Generator network and a Discriminator network. During training the Generator tries to fool the Discriminator into believing its output is real, while the Discriminator has to correctly spot the fake examples for the real ones. As training goes on the two networks become more and more accurate. GANs output is more sharp and defined that those of VAEs, but training is very unstable and computationally intense.

Such an approach has been used for music generation in MidiNet [21] and Musegan [22], showing promising results. Both models are based on CNNs, but while MidiNet uses a fairly classical CNN architecture, MuseGan uses an unprecedented model with a strong hierarchy that is based not on the single not but on the bar as a unit, featuring a bar generator controlled by a phrase generator. Both models can be conditioned to a priming sequence and MidiNet can also follow a chord progression. MuseGan is especially good for its ability to generate 5 instrument arrangements that sound rough, but show a degree of coherence. It has to be noted that GANs are still affected by the issues of LSTM and CNN if those architecture are part of the generator and discriminator networks.

E. *Constrained Markov Models*

A completely different approach relies on more classical machine learning. Constrained Markov Models used in [23] are a combination of Hidden Markov Models with Constrain Programming. Music is thus seen a Constraint Satisfaction Problem where Markov Models are responsible for generating music which follows a certain style learned from the data and CP is used to tackle the shortcomings of the Markov hypothesis applied to music, that is to enforce longer time dependencies. CP is also useful to avoid problems like repeated notes, as it simplifies the generation of music that follows a chord progression or a certain rhythmic structure.

This approach is used by the Flow Machines [24] and seems to perform particularly well in tasks that involve adapting a piece to the style taken from another. However the closed nature of the project makes it hard to tell where the contribution of the machine ends and that of the human begins.

III. New Interactive Tools

Google Magenta and the Flow Machine project by Sony CSL are two of the best research projects which focus on music generation and have produced some nice tools based on AI that help musician be more creative.

A. *Google Magenta*

Google Magenta[25] is an open source project started in 2017 with the goal of using AI to develop tools for creativity; most of their work is focused on music. Their first models were focused on generating melodies in response to a primer, this is basically the same as a language model used to suggest the next word.

With AI duet[26] they showcased their technology creating an AI ensemble that responded to the user input. While not perfect this showed the potential of deep learning for music. They continued developing open models for melody as well as polyphony generation. Their latest model is called MusicVAE[19] and has been used to create a number of different tools: Melody Mixer and Beat Blender allow to generate the interpolation between two melodies or drumbeats, while Latent Loops creates a grid with the 2-dimensional space between four melodies and allows the user to draw a path in the space that leads to the creation of a complex melody.

They also developed NSynth[27], a model that uses the same concept of latent space but applies it to timber instead of melodies. NSynth Super is a physical device that serves as an interface to this: the user puts 4 instruments on the 4 corners of a touchscreen and by swiping their finger in this 2-dimensional space they can select the interpolated sample. The interpolation being "semantic" generates sound that are not the simple superposition of the original sound but something that has the characteristics of all of them without being neither.

B. Flow Machines

The Flow Machines project [24] is led by Francois Pachet at Sony Computer Science Laboratories (Sony CSL Paris) and Pierre and Marie Curie University (UPMC). The goal of the project is to create tools for musicians. They gained popularity in 2016 with the release of "Daddy's Car" a track in style of the Beatles generated by their software; while the press described as the rise of the AI composers the truth is that the model generated the sheet music, in a style that sound like a mishmash of all the Beatles different styles, and this was then arranged and performed by the musician Benoît Carré. The song was created using Flow Composer, a tool that aids the composer by generating new melodies or by completing existing ones; the user can choose what parts he likes and what parts he wants changed or he can load an incomplete piece and how the AI fills the gaps. Flow Harmonizer has the ability to perform style transfer on an existing piece, examples show Ode to Joy in the style of Beatles, Bossa Nova, Bach, Jazz, etc.

Their Reflexive Looper [28] enhances the classic looper with AI that is aware of the style of the music being recorded as well as the instrument. It's then able to playback those loops accordingly as well as filling the missing pieces of the performance such as bass lines, guitar accompaniment and choirs. To do this it needs a classifier that can recognize the musical instrument being played and one for the musical style as well as a generator for that style that is also able to learn the style of the performer. This is achieved by considering both audio and video input of the performer.

IV. Market Perspective

In the last few years a number of startups have emerged that use AI to create music. Jukedeck and Amper Music focus on the creation of royalty-free soundtracks for content creator such as video-makers. Their service has tools to shape the music duration, style and climax to fit the specific video or presentation it's being created for. While the music create many not be very compelling the result is pleasing enough for a short soundtrack and the real value is in lifting the burden of copyright infringement from small creators.

Hexachords with his Orb Composer offers a more complex environment to create music that's not necessarily bound to be a soundtrack for a short video and it's more targeted to music professionals. The program does not replace the composer but rather assist him in an intelligent way, providing musical ideas that can be tweaked.

Luxembourg based Aiva Technologies focused on the creation an AI composer that produces scores to be recorded by real orchestras but it is not clear how this is achieved and to which degree human intervention is necessary.

These are just a few of a myriad of startups that are venturing in this uncharted territory but the common denominator of all of those companies seems to be the goal of enabling non-musical people to be creative and to enable musicians to be more effective and give them new path to follow in their endeavor.

V. Conclusion

Artificial intelligence can create new ways of interaction with music for composers, giving them the possibility to experiment and explore new frontiers. Smart instruments can also help people lacking a musical education to enjoy music composition, empowering them with tools that let them integrate their own music with other works. In this very challenging and exciting scenario, we sustain the view that research should focus on developing new tools to aid musicians rather than try to replace them. In this work we provided a brief overview of such a scenario, providing a technical discussion regarding the opportunities and problems that deep learning researchers are experiencing to this date.

References

[1] M. Roccetti, M. Gerla, C. E. Palazzi, S. Ferretti, and G. Pau, "First responders' crystal ball: How to scry the emergency from a remote vehicle," in *IEEE International Performance, Computing, and Communications Conference, 2007. IPCCC 2007.* IEEE, 2007, pp. 556–561.

[2] M. Roccetti, C. Prandi, P. Salomoni, and G. Marfia, "Unleashing the true potential of social networks: confirming infliximab medical trials through Facebook posts," *Network Modeling and Analysis in Health Informatics and Bioinformatics*, vol. 5, no. 1, 2016.

[3] M. Roccetti, G. Marfia, P. Salomoni, C. Prandi, R. M. Zagari, F. L. Gningaye Kengni, F. Bazzoli, and M. Montagnani, "Attitudes of Crohn's Disease Patients: Infodemiology Case Study and Sentiment Analysis of Facebook and Twitter Posts," *JMIR Public Health and Surveillance*, vol. 3, no. 3, p. e51, 2017. [Online]. Available: http://publichealth.jmir.org/2017/3/e51/

[4] A. Bujari, B. Licar, and C. E. Palazzi, "Road crossing recognition through smartphone's accelerometer," in *Wireless Days (WD), 2011 IFIP*. IEEE, 2011, pp. 1–3.

[5] A. Bujari, A. Marin, C. E. Palazzi, and S. Rossi, "Analysis of ECN/RED and SAP-LAW with simultaneous TCP and UDP traffic," *Computer Networks*, vol. 108, pp. 160–170, 2016.

[6] C. E. Palazzi, A. Bujari, S. Bonetta, G. Marfia, M. Roccetti, and A. Amoroso, "MDTN: Mobile delay/disruption tolerant network," in *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, 2011.

[7] S. Hochreiter and J. Urgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: http://www7.informatik.tu-muenchen.de/ hochreit%5Cnhttp://www.idsia.ch/ juergen

[8] T. Mikolov, M. Karafiát, L. Burget, J. Černock\'y, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3156–3164.

[11] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[12] D. Eck and J. Schmidhuber, "Finding temporal structure in music: Blues improvisation with LSTM recurrent networks," in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*. IEEE, 2002, pp. 747–756.

[13] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," *arXiv preprint arXiv:1206.6392*, 2012.

[14] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a Steerable Model for Bach chorales generation," *arXiv preprint arXiv:1612.01010*, 2016.

[15] B. L. Sturm, J. F. Santos, O. Ben-Tal, and I. Korshunova, "Music transcription modelling and composition using deep learning," *arXiv preprint arXiv:1604.08723*, 2016.

[16] Simon, Ian and Oore, Sageev, "Performance RNN: Generating Music with Expressive Timing and Dynamics," 2017. [Online]. Available: https://magenta.tensorflow.org/performance-rnn

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[19] A. Roberts, J. Engel, and D. Eck, "Hierarchical variational autoencoders for music," *NIPS Workshop on Machine Learning for Creativity and Design*, 2017.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[21] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR'2017), Suzhou, China*, 2017.

[22] H.-W. Dong, W.-Y. Hsiao, L.-C. Yang, and Y.-H. Yang, "MuseGAN: Symbolic-domain music generation and accompaniment with multi-track sequential generative adversarial networks," *arXiv preprint arXiv:1709.06298*, 2017.

[23] F. Pachet and P. Roy, "Markov constraints: steerable generation of Markov sequences," *Constraints*, vol. 16, no. 2, pp. 148–172, 2011.

[24] Sony Computer Science Lab, "Flow Machines." [Online]. Available: http://www.flow-machines.com/

[25] Google, "Magenta - Make Music and Art Using Machine Learning." [Online]. Available: https://magenta.tensorflow.org/

[26] Eck, Douglas, "Learning from A.I. Duet," 2017. [Online]. Available: https://magenta.tensorflow.org/2017/02/16/ai-duet

[27] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," *CoRR*, vol. abs/1704.01279, 2017. [Online]. Available: http://arxiv.org/abs/1704.01279

[28] M. Marchini, F. Pachet, and B. Carré, "Rethinking Reflexive Looper for structured pop music," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2017, pp. 139–144.